

GENETWORK



A database for post-genome analysis

When the Human Genome Project was initiated in the late 1980s, it was presented as the ultimate project to uncover the blue-print of life. Although the goal of sequencing the entire 30 billion base pairs of the human genome by 2005 is likely to be achieved, whether we will have the blue-print of life at that time is quite questionable.

First of all, as we have learned from the complete genomes of yeast and several bacteria, the biological function of a large fraction of the genes (a third to over a half depending on the organism) is still uncharacterized. Secondly and more importantly, given only the individual components that make up a biological system, the understanding of how each component works is not sufficient to understand the entire system. The post-genomic analysis, as we define it here, includes both experimental and informatics approaches to understand systems in terms of their integrated pathways of genes and molecules, which can be considered as the wiring diagrams of the biological system. The complete catalog of components and the complete catalog of wiring diagrams together can be called the *biogenetic map*.

KEGG (Kyoto Encyclopedia of Genes and Genomes) is an informatics project for the post-genome analysis, which we initiated in 1993 under the Human Genome Program of the Ministry of Education, Science, Sports and Culture in Japan. Its objectives are threefold. (1) To computerize the current knowledge of molecular pathways

and genetic pathways from the experimental observation of genetics, biochemistry, and molecular and cellular biology. In the past two years, KEGG contained only the metabolic pathways, but starting in July 1997 a number of regulatory pathways, such as signal transduction and gene development pathways, are being placed online. (2) KEGG maintains the gene catalog of every organism that has been sequenced, and each component in the catalog is to be mapped on to the KEGG pathways. (3) In addition to these database efforts, KEGG aims at developing new informatics tools for drug design associated with interactions and pathways.

KEGG is a part of the Japanese GenomeNet WWW server¹ and is linked to all the major molecular biology databases by the DDBJ LinkDB system². Figure 1 shows a portion of the phenylalanine, tyrosine and tryptophan biosynthesis, where each box represents an enzyme with the EC number inside. The box is clickable to retrieve the corresponding enzyme entry of the LIGAND database³, which is the starting point of retrieving related entries of chemical compounds, molecular sequences,

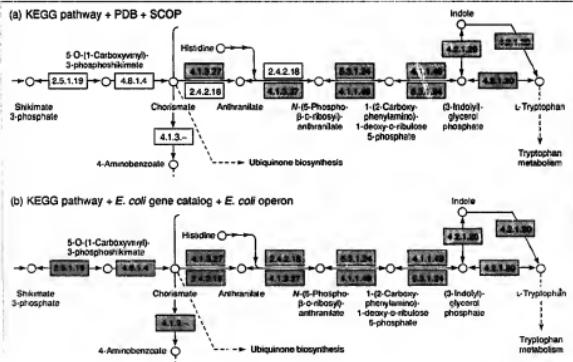


Figure 1. Examples of using KEGG as a large-scale genomic map tool. (a) From the KEGG table of contents choose the 'molecular catalog' feature. By selecting 'Pathway' select alpha beta 1 (the first box in the list). In the description in the category 'alpha beta 1 (Homo)', select the 'Pathway' option, paste in the search box, and search against '3D structures in PDB'. Several pathways, each of which contains at least one enzyme with a TM hand according to SCOP, will be indicated in the results screen. The view shown is one of phenylalanine, tyrosine and tryptophan biosynthesis. The blue boxes are the enzymes whose three-dimensional structures are known and those marked with red are the ones found to correspond to the SCOP classification. (b) From the KEGG table of contents choose the catalog of *Escherichia coli* operons, copy all the genes in the description in the 'tpo' operon, select the 'Pathway' option, paste in the search box, and search against 'Escherichia coli' operons. Select the 'vector' of phenylalanine, tyrosine and tryptophan biosynthesis pathway. The green boxes are the enzymes whose genes exist and those marked with red are the ones found to correspond to the 'tpo' operon.

GENETWORK

three-dimensional structures, and genetic diseases among others. KEGG maintains structural and functional classifications of molecules and genes in the form of what we call pathway diagrams. In some cases, linkages and subnetworks are clickable to unfold or fold branches. Figure 1(a) is a result (marked in red) of searching the *Ba* (Baumanni) pathway in the KEGG database, derived from the SCOP database⁴, with the KEGG metabolic pathway database⁵. In the pathway diagrams where the enzymes with known PDB codes are shown, the PDB codes are shown in blue boxes. This indicates possible gene duplications in the formation of the trypanosomal biosynthetic pathways⁶.

One of the most interesting aspects of KEGG is the automatic generation of organism-specific pathways by matching the gene cataloge produced by the genome sequencing project to the reference pathways that manually drawn and updated. In Figure 1(b) the enzymes colored in green indicate that the corresponding genes are found in the *Escherichia coli* genome. Enzymes colored in red are found belonging to the trypanosomal species, and the genome map section of KEGG can also be utilized with a Java-compatible browser, for example, to see any correlation between the physical proximity of genes

in the genome and the functional proximity of gene products in the pathway. An important consequence of mapping gene products on the pathway diagrams is the identification of missing gene products. In case the pathway is not continuous because of missing gene products, KEGG provides computational tools to assist researchers in the manual assessment and further analysis of possible existence of alternative paths⁷.

While KEGG sites to serve a large number of people at the level of abstraction are there, complementary resources

that contain more detailed data and knowledge in specific pathways. We have started discussions with WIT to generate

metabolic pathways that are open to other collaboration. The mirror sites of KEGG are being established⁸ in the USA and UK (Ref. 9). We plan to release a CD-ROM for Macintosh and Windows where pathway diagrams, hierarchical trees, and genome maps are all to be handled by a graphical user interface. The KEGG CD-ROM can be downloaded by anonymous FTP (Ref. 11) and used in UNIX as well. We also plan to start distributing the KEGG server to be mirrored in a local environment.

Minoru Kanehisa

kanehisa@kuicr.kyoto-u.ac.jp
http://www.kuicr.kyoto-u.ac.jp/~kanehisa/

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan

References

- 1 Goto, S. et al. (1996) Pacific Symposium on Biocomputing 1997, pp. 175–186.
- 2 <http://www.genome.ad.jp>
- 3 <http://www.genome.ad.jp/dbget/>
- 4 http://www.genome.ad.jp/html/show_main.html
- 5 Moerlin, A.C., Brenner, S.E., Hwang, T. and Chothia, C. (1995) *J. Mol. Biol.* 253, 549–559.
- 6 Wilenmann, M. and Eisenberg, D. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 1379–1383.
- 7 <http://www.genome.ad.jp/kegg/comp/comp.html>
- 8 <http://www.genome.ad.jp/kegg/comp/comp.html>
- 9 <http://www.cme.msu.edu/wt/>
- 10 <http://kegg.genome.ad.jp/CD/>

TECHNICAL TIPS ONLINE

<http://www.elsevier.com/locate/tio> [\(In Europe\)](http://www.elsevier.nl/locate/tio)

Technical Tips Online publishes short, peer-reviewed, molecular biology techniques articles in a Web-based environment. The articles describe novel methods or significant improvements to existing methods in any aspect of molecular biology.

New Technical Tip articles published recently in Technical Tips Online include:

Gao, C., Nickle, D.G. and Pritchett, P. (1997) A PCR method for determining the presence of multiple (12) and C57BL/6 mice gene knockout strains. Technical Tip Online (<http://www.elsevier.com/locate/tio/10119>)

Granger, R.L. (1997) A glass bead method for picking bacterial colonies. Technical Tip Online (<http://www.elsevier.com/locate/tio/10117>)

Herrlich, S., Mazzoni, F., Lerninger, C. and Boulton, J. (1997) There is still life after death for yeast Kef1 antibodies. Technical Tip Online (<http://www.elsevier.com/locate/tio/10116>)

Lau, D. and Gackstetter, N. (1997) Use of a simplified and rapid slot-sieve protocol for the detection of recombinant plasmids. Technical Tip Online (<http://www.elsevier.com/locate/tio/10112>)

Luck, A., Im, Y.H., Sengupta, P. and Massie, B. (1997) A method for preparing DM-labeled thiotores that allows single-cell resolution with conventional fluorescence microscopy. Technical Tip Online (<http://www.elsevier.com/locate/tio/10119>)

Morochek, J. (1997) Two yeast/*Escherichia coli* plasmid vectors designed for yeast one- and two-hybrid screens that allow directional cDNA cloning (<http://www.elsevier.com/locate/tio/10111>)

Patterson, N., Becker, J., Gordon, M. and Rev, M. (1997) An inverse PCR strategy for the recovery of full fragments from the phage λ C2 vector. Technical Tip Online (<http://www.elsevier.com/locate/tio/10116>)

Schramm, H. (1997) Rapid high-throughput purification of genomic DNA from mouse and rat tails for use in transgenic testing. Technical Tip Online (<http://www.elsevier.com/locate/tio/10115>)

Spiekerman, F., Van Gele, M., Maertens, L. and Van Roy, N. (1997) Improved protocol for chromatofocusing from fixed cells (<http://www.elsevier.com/locate/tio/10113>)

Spronghuis, B., Heng, H.J.U.Q. and Meers, P.H. (1997) Recycling cells in IEM and immunocytopathology studies. Technical Tip Online (<http://www.elsevier.com/locate/tio/10109>)

Editor Adrian Bird, Institute for Cell and Molecular Biology at the University of Edinburgh